Assessment of Hydration Thermodynamics at Protein Interfaces with Grid Cell Theory

Assessment of Hydration Thermodynamics at Protein Interfaces with Grid Cell Theory

Georgios Gerogiokas[a], *Michelle W. Y. Southey*[b], *Michael P. Mazanetz*[b], *Alexander Heifetz*[b], *Michael Bodkin*[b], *Richard J. Law*[b], *Richard H. Henchman*[c], *and J. Michel*[a*]

[a]EaStCHEM School of Chemistry, Joseph Black Building, The King's Buildings, Edinburgh, EH9 3JJ, UK. E-mail: mail@julienmichel.net

[b]Evotec (UK) Limited, 114 Innovation Drive, Milton Park, Abingdon, Oxfordshire OX14 4SA.

[c]Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, United Kingdom and School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

Abstract: Molecular dynamics simulations have been analyzed with the Grid Cell Theory (GCT) method to spatially resolve the binding enthalpies and entropies of water molecules at the interface of 17 structurally diverse proteins. Correlations between computed energetics and structural descriptors have been sought to facilitate the development of simple models of protein hydration. Little correlation was found between GCT computed binding enthalpies and continuum electrostatics calculations. A simple count of contacts with functional groups in charged amino-acids correlates well with enhanced water stabilization, but the stability of water near hydrophobic and polar residues depends markedly on its coordination environment. The positions of X-ray resolved water molecules correlate with computed high density hydration sites, but many unresolved waters are significantly stabilized at the protein surfaces. A defining characteristic of ligand-binding pockets compared to non-binding pockets was a greater solvent-accessible volume, but average water thermodynamic properties were not distinctive from other interfacial regions. Interfacial water molecules are frequently stabilized by enthalpy and destabilized entropy with respect to bulk, but counter-examples occasionally occur. Overall detailed inspection of the local coordinating environment appears necessary to gauge thermodynamic stability of water in protein structures.

1. Introduction

Water plays a crucial role in the structure and dynamics of proteins. Water has been implicated as a mediator of interactions between different protein surfaces,[1] and is a major driver for protein folding through the burial of hydrophobic side chains of amino acids.[2] Understanding water-protein interactions relates to protein function and is important for enzyme catalysis, as well as DNA-water interactions[3] and molecular recognition of various events including protein-DNA,[4,5] protein-protein[6] and protein-ligand interactions.[7] Greater understanding of the role played by water in molecular recognition opens up new avenues for the creation of novel therapeutics.

A key question relates to the thermodynamics properties of water at the interface of biomolecules, sometimes also called biological water.[8] Biological water is often defined as a hydration layer around proteins. This hydration layer is distinct from bulk water both thermodynamically and dynamically, as shown from terahertz spectroscopy data,[9] and molecular dynamics.[10] It is important to understand the extent of this hydration layer and whether the majority of cellular water does differ greatly from bulk.[10] Most experimental and molecular dynamics studies suggest only the first two solvation shells significantly differ from bulk water when the oxygen density of water molecules is considered, but orientational correlations may be longer-ranged.[11,12] For water in the first hydration layer of biomolecules there is clear coupling between dynamics and thermodynamics.[13]

The present study is primarily concerned with the correlation of interfacial water thermodynamics with protein structural descriptors. The dataset includes approximately 85,000 hydration sites across the interface of 17 proteins. Many of these proteins are popular drug targets such as: HMG-COA reductase, PDE5, cyclooxygenase, caspase1, MDM2, kinases (CDK,

cAbl), thrombin, HIV, neuraminidase, penicilin binding protein. This dataset overlaps with a dataset used by Beuming *et al.*[14] to evaluate thermodynamics of hydration sites using the inhomogeneous fluid solvation theory (IFST) as implemented in the *Watermap* software.[15] A secondary objective of the present study was thus to compare IFST computed water thermodynamic properties to those produced by the Grid Cell Theory (GCT) methodology, as implemented in the software *Nautilus*. GCT is a newly developed method to investigate hydration thermodynamics from a single molecular dynamics (MD) or Monte Carlo (MC) trajectory. GCT is a spatial discretization of the cell theory method developed by Henchman.[16] GCT has recently been validated on small molecules,[17] model binding sites,[18] as well as protein-ligand complexes.[19] Other novel analyses that are reported here include correlation of Poisson-Boltzmann electrostatics with water binding thermodynamics, water thermodynamics in binding-sites, and correlations between water binding enthalpies and entropies. The results help build a comprehensive picture of the hydration thermodynamics at protein interfaces, and suggest how complexities may be subsumed into simpler structural descriptors.

2. Theory and Methods

Grid cell theory

Grid cell theory has been used to compute binding free energies as reported in previous work.[17–19] In the approach outlined the density, enthalpy, entropy and free energy of water are evaluated for an arbitrary region of space ***s*** around a system of interest *X*. Binding free energy, enthalpy and entropy of water defined here refer to the process where water enters a particular hydration site of the protein(s) from bulk concentration. Here the computation of the binding free energy involves three steps.

First, parameters of water molecules inside ***s*** are evaluated. For each frame *f*, cell parameters of the N_f water molecules $i \in \boldsymbol{s}$ are determined. These cell parameters are: the magnitude of the components of the intermolecular forces $|F_i^j|$ and torques $|\tau_i^j|$ along the principal axes *j* (*j* = *x,y,z*) of the water molecule *i*, the orientational number Ω_i^{ori} of the water molecule *i*, the protein-water interaction energy ΔH_i^X, and the water-water interaction energy ΔH_i^w In line with preceding studies, the water-water interaction energy term is half the average interaction energy with other water molecules, minus half the average interaction energy in bulk water. These quantities are equated to enthalpies because contributions from pressure-volume terms were neglected. Detailed expressions for these quantities may be found elsewhere.[19]

Second, parameters for volume elements within ***s*** are determined. To do so the region ***s*** is decomposed into N_s voxels of volume *V(k)*. Properties of each *k* voxel are given by equation 1:

$$A(k) = \frac{\sum_{f=1}^{M} \sum_{i=1}^{N_f} A_i \, I_k(i)}{max\left\{1, \sum_{f=1}^{M} \sum_{i=1}^{N_f} I_k(i)\right\}}, \quad (1)$$

where typically $A_i = F_i^j$, τ_i^j, and $\Omega_i^{ori}, \Delta H_i^X$, and, ΔH_i^w. $I_k(i)$ is an indicator function which is equal to 1 if water molecule *i* is in voxel *k*, and 0 otherwise. Whether a water molecule *i* is within a voxel is determined by inspection of the Cartesian coordinates of the oxygen atom of the water molecule. Finally, *M* is the number of frames in the analyzed trajectory. The average number of water molecules within voxel *k* is given by equation 2:

$$N_w(k) = \frac{1}{M} \sum_{f=1}^{M} \sum_{i=1}^{N_f} I_k(i) \quad (2)$$

Third, binding thermodynamic properties of ***s*** are evaluated. Equations 3 and 4 give the solute and solvent components of the enthalpy of binding of region ***s***:

$$\Delta H_X^{\boldsymbol{s}} = \sum_{k=1}^{N_s} N_w(k) \, \Delta H_X \, (k) \quad (3)$$

$$\Delta H_w^{\boldsymbol{s}} = \sum_{k=1}^{N_s} N_w(k) \, \Delta H_w \, (k) \quad (4)$$

The enthalpy of binding in region $\boldsymbol{s}$, is given by equation 5:

$$\Delta H^{\boldsymbol{s}}_{w,X} = \Delta H^{\boldsymbol{s}}_{X} + \Delta H^{\boldsymbol{s}}_{w} \quad (5)$$

The average number of water molecules within $\boldsymbol{s}$ is computed with equation 6:

$$N_w(\boldsymbol{s}) = \sum_{k=1}^{N_s} N_w(k) \quad (6)$$

The average orientational numbers and forces/torques for region $\boldsymbol{s}$, are given by equation 7:

$$A(\boldsymbol{s}) = \frac{1}{N_w(\boldsymbol{s})} \sum_{k=1}^{N_s} N_w(k)\, A(k), \quad (7)$$

where $A = F^j$, τ^j, and Ω^{ori}. Additionally, the minimum value for $\Omega^{ori}(\boldsymbol{s})$ is always 1 in this work. The calculation of the orientational number of each water molecule *i* in frame *f* is based on the generalized Pauling's residual ice entropy model given by equation 8, where N_a is the number of hydrogen bond acceptors within 3.4 Å around water *i*. This equation is used unless there is a solute polar or charged atom in the coordination shell of the water, in which case equation 9 is applied instead.

$$\Omega^{ori} = \frac{N_a(N_a-1)}{2} \left\{ \frac{N_a-2}{N_a} \right\}^2, \quad (8)$$

$$\Omega^{ori} = \frac{N_a^{eff}(N_a^{eff}-1)}{2} \left\{ \frac{N_a^{bulk}-2}{N_a^{bulk}} \right\}^{2-p_{HB}^X}, \quad (9)$$

where N_a^{eff} is the effective coordination number, N_a^{bulk} is the coordination number of bulk water. The number of hydrogen bond acceptors N_a is given by:

$$N_a = N_X + N_{w_s} + N_{w_b} \quad (10)$$

where N_x is the number of solute acceptor atoms within the cutoff, N_{ws} is the number of first hydration shell water molecules within the cutoff, and N_{wb} the number of remaining water molecules. Next, the ratios of each type of acceptors that are hydrogen bonded to water *i* is given by equation 11.

$$p_{HB}^{X} = \frac{N_{XHB}}{N_X} \; ; \; p_{HB}^{w_s} = \frac{N_{w_sHB}}{N_{w_s}} \; ; \; p_{HB}^{w_b} = \frac{N_{w_bHB}}{N_{w_b}} \tag{11}$$

And the effective coordination number is then obtained from equation 12:

$$N_a^{eff} = \frac{N_{XHB}+N_{w_sHB}+N_{w_bHB}}{\max(p_{HB}^{X},p_{HB}^{w_s},p_{HB}^{w_b})} \tag{12}$$

With the orientations, forces, and torques equations 13-14 are used to give the entropic components:

$$\Delta S_{w,X}^{\boldsymbol{s},ori} = N_w(\boldsymbol{s}) k_B \ln\left\{\frac{\Omega^{ori}(\boldsymbol{s})}{\Omega^{ori}(bulk)}\right\} \tag{13}$$

$$\Delta S_{w,X}^{\boldsymbol{s},vib} = N_w(\boldsymbol{s}) k_B \ln\left\{\prod_{j=1}^{3} \frac{F^j(bulk)}{F^j(\boldsymbol{s})}\right\} \tag{14}$$

$$\Delta S_{w,X}^{\boldsymbol{s},lib} = N_w(\boldsymbol{s}) k_B \ln\left\{\prod_{j=1}^{3} \frac{\tau^j(bulk)}{\tau^j(\boldsymbol{s})}\right\} \tag{15}$$

where $\Omega^{ori}(bulk)$, $F^j(bulk)$, $\tau^j(bulk)$ (j = *x, y, z*) are the cell parameters for the simulated water model in bulk conditions. Summation of the components in equation 16 allows the computation of the entropy of binding within region ***s***:

$$\Delta S_{w,X}^{\boldsymbol{s}} = \Delta S_{w,X}^{\boldsymbol{s},ori} + \Delta S_{w,X}^{\boldsymbol{s},vib} + \Delta S_{w,X}^{\boldsymbol{s},lib} \tag{16}$$

Finally, the addition of the enthalpic and entropic components gives the binding free energy of water within region **s**:

$$\Delta G_{w,X}^{\boldsymbol{s}} = \Delta H_{w,X}^{\boldsymbol{s}} - \mathrm{T}\Delta S_{w,X}^{\boldsymbol{s}} \tag{17}$$

Nautilus is a trajectory analysis software that implements equations 1-17. For some analyses, the enthalpy, entropy and free energy values of a region **s** were further normalized by number of waters present within region **s**, and this is denoted by the superscript symbol '**w**'.

Nautilus, has several dependencies including the molecular simulation framework *Sire*,[20] and the *MDtraj* python package.[21] Molecular models used, regions chosen and molecular simulation protocols are further described below.

Preparation of molecular models

The following 17 PDB[22] structures were used: 4COX[23], 1BMQ[24], 1E1X[25], 1E9X[26], 1E66[27], 1EZQ[28], 1HWL[29], 1HWR[30], 1IEP[31], 1KV1[32], 1M17[33], 1NLJ[34], 1OYN[35], 1PTY[36], 1QMF[37], 1UDT[38], and 1YCR[39]. The structures were obtained from the initial PDB structure after the respective ligands and/or co-solutes were removed (if there were dimers or homodimers only the relevant monomer was used). After any ligands were removed, the software *tleap* (Amber 11)[40] was used to parameterize the system using the AMBER99SB forcefield,[41] and the TIP4P-EW water models for the solvent.[42] The system was solvated in a rectangular box whose edges extended at least 11 Å away from the edges of the protein. Where appropriate and in agreement with the experimental data, cysteine pairs were modelled as disulfide bonds. Each solvated protein was first energy minimized and then equilibrated with positional restraints for 1 ns with the *sander* module before production runs.

Molecular dynamics simulation protocols

All subsequent molecular simulations were produced using the software *Sire*/*OpenMM* (*SOMD*). In this study the software *SOMD* results from the linking of the general purpose molecular simulation package Sire (revision 1786)[20], with the GPU molecular dynamics library *OpenMM* (revision 3537)[43]. Simulations were run at a pressure of 1 atm and temperature of 298 K using an atom-based Barker-Watts reaction field non-bonded cutoff of 10 Å for the electrostatic interactions with a dielectric constant set to 78.3,[44] and an atom-based non-bonded cutoff of 10 Å for the Lennard-Jones interactions. A velocity-Verlet integrator with a time step of

2 fs was used. Temperature control was achieved with an Andersen thermostat with a coupling constant of 10 ps^{-1}.[45] Pressure control was implemented via attempted isotropic box edge scaling Monte Carlo moves every 25 time steps. The *OpenMM* default error tolerance settings were used to constrain the intramolecular degrees of freedom of water molecules. For each protein system one simulation of 50 ns was run with velocities randomly drawn from a Maxwell-Boltzmann distribution. In all simulations a harmonic restraint $\boldsymbol{r}_p$ was placed on all heavy atoms of a protein P with a force constant of 10 kcal mol^{-1} $Å^{-2}$ and reference coordinates taken from the initial structure. The use of restraints influences the computed water thermodynamics and this is made explicit in the notations used in the rest of this manuscript by use of the subscript '$P(\boldsymbol{r}_p)$' instead of 'X' in the notations defining computed thermodynamic quantities. Snapshots were saved every 1 ps in a DCD format. The first 1 ns of equilibration was not included in the data averaging.

Nautilus analyses

The *Nautilus* post-processing tool was used to generate rectangular grids around the protein. For each protein the grid was placed so that it extends at least 4.0 Å away from the extreme edges of each protein at 1 Å grid density. This cutoff was deemed sufficient to analyses first hydration shell interfacial waters. In *Nautilus* the grid is defined by specifying a coordinate center (x_c,y_c,z_c) and from this center specifying the maximum and minimum grid positions as ($x_c \pm \Delta x$, $y_c \pm \Delta y$,$z_c \pm \Delta z$). Since all heavy atoms in the protein are restrained, spatial variations in hydration thermodynamics are captured on the 3D grid via simple averaging of the MD trajectories. Various regions around amino-acid side chains, clustered sites, or predicted pockets are also chosen as seen in figure 1. Water binding free energies for these regions are then computed and correlated to other descriptors reported below.

Amino-acid environment analyses

GCT properties were computed for each type of amino acid from the dataset of 17 proteins. A distance cutoff of 4 Å was used to select a set of grid points near specific amino acids throughout the dataset. Beuming et al.[14] used a similar cutoff in their IFST study, but there the cutoff was only used to bin density clustered hydration sites. GCT does not need an a priori definition of hydration sites and water properties over all grid points within the specified cutoff are considered. One advantage is that water behavior is resolved even in spatial regions of low solvent density.

The cutoff was applied to select grid points near functional groups rather than entire amino acids. The different groups were: *carboxylic acids* (aspartates and glutamates), *side-chain nitrogen(s)* (lysines and arginines), *hydroxyl groups* (threonine, serine, tyrosine), *side-chain amides* (glutamine and asparagine), *ring atoms* (tyrosine, histidine, phenylalanine and tryptophan), *non-polar side-chains* (leucine, isoleucine, valine, and alanine) including hydrogens in the side-chains.

The resulting distributions of water thermodynamic properties for each group were then compared using a Kolmogorov-Smirnov test as provided by the R programming language.[46] This nonparametric statistic measures the likelihood that two sets of samples were derived from the same underlying distribution. The empirical cumulative distribution function $F_n(x)$ is given by equation 13:

$$\mathrm{F}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\infty,x]}(X_i), \tag{13}$$

where n observations have been binned by the indicator function $I_{[-\infty,x]}$ which is equal to 1 if $X_i \leq x$ otherwise it is equal to zero. This procedure is repeated for both datasets and then a Kolmogorov-Smirnov statistic D_n is computed with equation 14:

$$D_n = sup_x|F_n(x) - \mathrm{F}(\mathrm{x})|, \tag{14}$$

where sup_x is the supremum, or lowest upper bound of the set of distances derived from the two empirical cumulative distribution functions. A D_n value of zero signifies no difference in the distribution, whereas a value greater than ca. 0.2 for the datasets analysed here suggests a low probability (p-value < 0.05) that the samples are drawn from the same distribution. P-values between distributions are shown in the supplementary information (fig S1).

Density-clustering of hydration sites

Density clustered sites were calculated by assigning to a cluster the grid point with the highest density. This is not necessary to compute thermodynamic properties with GCT, but useful to analyse regions of high water density. All grid points within a neighbor cutoff of that grid point are then assigned to that same cluster. This procedure is then iterated until all grid points have been assigned to a cluster or until no points with a density above a threshold value remain. Here the neighbor cutoff was set at 1.5 Å (roughly the radius of a water molecule) and a density threshold of at least 1.5× that of bulk water.

Analysis of crystallographic hydration sites

Hydration sites obtained via density-clustering of MD trajectories were compared with hydration sites observed in X-ray diffracted crystal structures with the following protocol. First, each PDB protein structure (including hydration sites) was aligned on to the simulation frame of reference using all heavy atom backbone atoms. Then a density-clustered grid was produced to obtain clustered sites from simulation data. Finally, for each experimental hydration site the minimum distance to a density-clustered site was calculated.

Comparison of pockets and binding sites

In this analysis the hydration thermodynamic properties of up to the top 10 druggable pockets of a protein structure as predicted by the software *fpocket*[47] are compared to those of the known

binding site. This software detects pockets in proteins by using alpha spheres.[48] Alpha spheres are defined as spheres which must contact at least 4 atoms within a cut-off distance from the alpha sphere center. These alpha spheres in turn reflect the local curvature: in a protein, buried pockets tend to be occupied by larger quantities of small radii alpha spheres, the surface is typically composed of larger radii alpha spheres, and intermediate radii usually reflect more exposed binding sites and clefts.

The general workflow for this analysis is: 1) Use *fpocket* to generate up to top 10 druggable pockets for each protein in the dataset. 2) Extract pocket coordinates from *fpocket* output. 3) Select all *Nautilus* grid points within 1 Å of any pocket-site coordinates to define a spatial region. 4) Binding thermodynamics for water in this region are computed by grouping cells in the region.

This protocol produces for each pocket per-site $\Delta G^{\mathbf{s}}_{w,P(\mathbf{r}_p)}$, $\Delta H^{\mathbf{s}}_{w,P(\mathbf{r}_p)}$,$-T\Delta S^{\mathbf{s}}_{w,P(\mathbf{r}_p)}$, values, as well as per-water $\Delta G^{\mathbf{s,w}}_{w,P(\mathbf{r}_p)}$, $\Delta H^{\mathbf{s,w}}_{w,P(\mathbf{r}_p)}$,$-T\Delta S^{\mathbf{s,w}}_{w,P(\mathbf{r}_p)}$ values, relative density, average number of waters, and the solvent-accessible volume of the pocket. For fifteen of the seventeen structures considered here pockets computed by *fpocket* overlapped well with the location of a known ligand binding site (one site in 4COX, 1E1X, 1E66, 1E9X, 1EZQ, 1IEP, 1KV1, 1M17, 1NLJ, 1OYN, 1UDT and two binding sites in 1PTY and 1QMF), and these sites were included in the analysis.

Comparison of electrostatic potential with binding enthalpies

The goal here was to establish whether the magnitude of the electrostatic potential at particular region of space correlates with the GCT computed enthalpies. To enable a reasonable comparison, only density-clustered sites within 4 Å of the protein obtained from simulations

were assessed with Poisson-Boltzmann calculations. This effectively discards regions of space that have high values of the electrostatic potential but are not solvent accessible for steric reasons.

The *APBS* Poisson-Boltzmann solver of Baker et al.[49] was used and the following protocol was used to implement this analysis: 1) Generate a large coarse grid with *APBS* but specify that the fine grid contains the same spacing and density used for the *Nautilus* grids; 2) High-density hydration sites are obtained from the GCT density-clustering method discussed previously in the section density-clustering of hydration sites; 3) The magnitude of the electrostatic potential of each *APBS* grid point that belongs to a given hydration site is computed and averaged; 4) The average magnitudes are compared with the GCT computed binding enthalpy of that hydration site.

Results and Discussion

Hydration thermodynamics near amino acids

The average free energies, enthalpies and entropies of binding of waters near each type of side chain are shown in figure 2. Kolmogorov-Smirnov (KS) tests were also used to estimate the likelihood that the observed distributions were drawn from the same underlying distribution. Figure 3A shows an example of a histogram of the sampled distribution for Alanine (histograms of other distributions are shown in the supplementary information figures S2-S3). Figure 3B shows a heatmap of D_n values. Finally differences in average properties between amino acid groups: polar, negatively charged, positively charged, aliphatic, and aromatic types of amino acids are shown in figure 4.

Three trends emerge from analysis of Figures 2-4: first, the majority of the free energy changes have a large enthalpic contribution in comparison to the entropy. The second trend observed is

that negatively charged side chains stabilize waters significantly more than any other amino acid type. Thirdly, polar, aliphatic and aromatic have similar variations to each other in the range of free energy values which suggest water stability in these regions should be assessed on a case-by-case basis.

Negatively charged amino acids show a clear stabilisation of waters with average binding free energies of −6.99±0.19, and −6.94±0.14 kcal mol^{-1} $water^{-1}$ for aspartate and glutamate respectively (figure 2). These two amino-acids both have similar distributions and this is reflected by the low D values of the KS tests (figure 3B). Next, amino acids that are positively charged were analyzed. Arginine and lysine have more similar free energy distributions while histidine seems to have a broader distribution, which seems to be related to the different protonation states which were grouped together for this analysis. Thus for histidine it is instructive to analyse the results for the delta/epsilon tautomers and the doubly protonated form. The average binding free energies are −4.82±0.94 (delta tautomer), −4.98±0.54 (epsilon tautomer), and −8.60±1.71 kcal mol^{-1} $water^{-1}$ (doubly protonated) respectively. Thus the particularly broad distribution for histidine is due to the charged tautomer. Polar amino acids were also analyzed and threonine stabilizes water the most, followed by serine and asparagine as shown in figures 2 and 3B. Interestingly, amino-acids containing an amide side-chain (asparagine, glutamine) stabilized waters less well than hydroxyl containing functional groups, and showed also greater differences their distributions according to the KS test shown in figure 3B. On the other hand, aliphatic amino acids show little difference between the free-energy distributions of alanine, leucine or isoleucine but larger variations are seen for methionine as shown in Figure 3B. This could be due to the effect of the sulfur atom on the sidechain that confers a different environment for solvating water molecules. Aromatic amino acids

(phenylalanine, tryptophan, and tyrosine) show few differences amongst themselves, which is also reflected by their similar average free energy per water shown in figure 2. Overall, water near negatively charged amino acids are stabilized the most, followed by water near positively charged amino acids. All other types of amino acids do not exhibit significantly different distributions.

Next *Watermap* values obtained from the Beuming et al. work[14] (figure 4) were compared to the present results. In the IFST study of Beuming et al[14] entropy changes are invariably unfavorable because the theory used assumes that bulk water is uniformly distributed and has thus no correlations, but water-protein interactions always introduce correlations that decrease water entropy. This assumption appears substantial, indeed Beuming et al.[14] observed ΔS entropy losses for water of up to 20.1 cal mol^{-1} K^{-1} (corresponding to $-T\Delta S = +6$ kcal.mol^{-1}) which exceeds the entropy of bulk water (16.7 cal $mol^{-1}K^{-1}$). Cell theory makes no such assumption because entropy changes are based upon changes in cell parameters calculated for water in the liquid state and at the interface of a protein. This yields entropy changes that are generally unfavorable but of smaller magnitude than those reported by Beuming et al.[14]

Average binding enthalpies appear to be more negative in the present GCT analyses than those reported by Beuming et al.[14]. This may be due to differences in the protocols used to detect and define hydration sites but it is difficult to be certain. The GCT binding enthalpies are in better agreement with those computed by Huggins using IFST for 23 hydration sites.[50] Huggins reported binding enthalpies ranging from −18.7 kcal mol^{-1} to −3.9 kcal mol^{-1}, with a mean value of −10.3 kcal mol^{-1}. The mean GCT binding enthalpies reported in Figure 4 are more positive, but this is likely because the present analyses include a very large number of interfacial hydration sites, whereas Huggins study focused on hydration sites found in internal protein

cavities. The minimum and maximum computed binding free energies per amino acid in this study (Figures S2-S3, typically −18 to 0 kcal.mol^{-1}) are indeed consistent with the range of binding free energies computed by Huggins. The range of GCT derived binding enthalpies is also similar to that computed with another IFST implementation in earlier work by Li and Lazaridis for a range of proteins (−19.2 to −1.4 kcal mol^{-1}).[51] Earlier FEP work from Hamelberg et al. reported binding free energies in the range of −0.8 to −3.4 kcal mol^{-1},[52] whereas Michel et al. reported FEP-derived standard binding free energies for water molecules observed in various X-ray structures in the range of ca. -4 to -11 kcal mol^{-1}.[53-54] Overall these figures are also consistent with the range of GCT computed binding free energies.

Finally, the negatively charged amino acids in GCT tend to decrease the entropy significantly more than any other amino acid type. However with the IFST implementation of Beuming et al. there is not as large a difference in the per-amino acids variations of the entropy of solvating water molecules. Despite these differences, the overall ranking of the amino acids with both methods follows similar trends.

Crystallographic water analysis

Next the positions of 1716 crystallographic water sites (derived from all proteins of the dataset except: PDBs 4COX, 1BMQ, 1HWR, 1NLJ and 1YCR which did not contain crystallographic waters) were compared to the position of clusters derived from grid densities computed from molecular dynamics snapshots. Figure 5 shows how the density of MD-derived hydration sites varies as a function of the minimum distance to a hydration site observed in an X-ray diffracted protein crystal. The figure shows that MD-derived hydration sites closer to the crystal water sites tend to have densities greater than bulk water densities, as expected. Conversely, hydration sites further away from a crystal site usually have more bulk-like water densities. There are more

sites with bulk-like density at minimum distances of ca. 15-20 Å from a crystallographic water site. These sites are typically not observed in X-ray diffracted structures. As expected this confirms that crystallographic techniques are better suited at discerning denser water sites, rather than low density water sites. Nevertheless, a sizable number of high density hydration sites are also observed far from any crystallographically resolved site.

Comparison of Poisson-Boltzmann electrostatic potentials with binding enthalpies

A comparison of the binding enthalpies of 29,507 high density sites (density greater than 1.5 times bulk density) within 4 Å of a protein with the computed magnitude of the electrostatic potential at each site is shown in Figure 6. The comparison is done with the average of the magnitude of the electrostatic potential because sites that contain several grid points with high positive or negative values of the electrostatic potential may have a signed average potential close to zero, which does not distinguish them from sites made of grid points with uniformly low values of the electrostatic potential. Comparison with binding free energies would give broadly similar results because variations in binding enthalpies are the dominant contribution to binding free energies (see Figure 9B below). Hydration sites which were found to have a high positive enthalpy of binding (above 2.6 kcal mol^{-1}) all contributing from $\Delta H^{\boldsymbol{s,w}}_{w,P(\boldsymbol{r}_p)}$. Further analyses indicated that these sites correspond to regions where a water molecule was sterically hindered and trapped, possibly due to the use of positional restraints on the protein atoms these were removed from subsequent analyses. In general, most sites tend to have a low average magnitude in their electrostatic potential, but this does not imply poor binding enthalpies as evidenced by the wide scatter of binding enthalpies seen in Figure 6. Indeed, any correlation between the enthalpy of binding and the magnitude of the electrostatic potential is very weak. Thus it appears that the stability of a water molecule may not be reliably determined from local values of the

electrostatic potential. This is illustrated with some examples taken from this dataset. Figure 7A displays one example with a low magnitude of the electrostatic potential, but large negative enthalpy of binding. The low magnitude of the local electrostatic potential (10.0 $k_BTe_c^{-1}$) occurs due to cancellation of electric fields induced by a nearby aspartate and two arginine side-chains. The enthalpy of binding is fairly negative (−18.1 kcal mol^{-1}) as a result of good coordination of an oxygen water with two arginine side chain nitrogens, an interaction with an aspartate sidechain oxygen with one of the water's hydrogens, and finally an interaction with a neighboring water molecule. Figure 7B shows a case of fairly negative enthalpy of binding (−21.8 kcal mol^{-1}) and high magnitude of the average local electrostatic potential (46.1 $k_BTe_c^{-1}$). The coordination environment of the hydration site is fairly similar to the site in Figure 7A, it involves one aspartate, two arginines and a threonine hydroxyl group. Consequently, the enthalpy of binding is similar, but the electrostatic potential differs significantly. Figure 7C shows an example where the electrostatic potential has high magnitude (59.3 $k_BTe_c^{-1}$). This occurs because the hydration site is close to a positively charged lysine. The electric field induced by this residue is not offset by neighboring negatively charged side-chains. However, the enthalpy of binding is poor (−1.8 kcal mol^{-1}) because water in this environment is unable to coordinate effectively with the lysine's ammonium group and can engage in at most one hydrogen-bond with a neighboring water molecule. Taken together Figure 6 and 7 shows that continuum electrostatic calculations may not be reliably used to estimate the stability of a hydration site, and inspection of the coordinating environment is a better indicator of thermodynamic stability.

Comparison of protein pockets with ligand binding sites

Figure 8A depicts the distributions of free energy and enthalpy of binding of water in known binding sites and other pockets found in the dataset. Enthalpy of binding is the largest component of the binding free energy of the pockets. After normalization with respect to the number of water molecules there is no significant difference between the distributions, given the available data (Figure 8A). The mean binding enthalpies are both -3.1 ± 0.2 kcal.mol^{-1}, and the standard deviations 0.9 and 2.2 kcal.mol^{-1} for the binding site and pocket datasets respectively. For the entropy of binding the per-water statistics are also comparable (Figure 8B). Finally, Figure 8C shows that binding sites contain a larger number of water molecules than pockets, and this is because the volume of the binding sites is larger (Figure 8D).

Taken together, these results suggest that these binding sites do not appear to generate an unusual signature in the computed hydration thermodynamics when average water properties over a complete pocket are considered. Rather the location of ligand binding sites in proteins may inferred by analysis of the solvent-accessible volume of pockets. This findings contrast with reports from Beuming et al.[14] or Vukovic et al.[55] that developed druggability descriptors based on computed water binding thermodynamics. The main differences with the present work are that these studies focused on detection of the least stable (or clusters of) high-density hydration sites, whereas here average per-water properties over a larger volume of space were considered.

Thermodynamic properties of high-density hydration sites

Next high-density hydration sites were analyzed further as these often involve structured water that are important for protein stability and/or function. Most sites have a free energy of binding between 0 to –15 kcal mol^{-1}, with extreme cases reaching up to –50 kcal mol^{-1} There is only a weak anti-correlation between the entropy and free energy (Figure 9A). Entropies of binding are

generally positive (up to ca. 2.5 kcal mol^{-1}) but in some instances negative entropies of binding are observed. Figure 9B shows that by contrast there is a strong correlation between the enthalpy and free energy and this reflects also the large contribution that this component makes to the free energy.

Hydration sites with unusual entropies of binding were further inspected. Figure 10A depicts a site with a tightly bound water molecule ($\Delta G^{s}_{w,P(r_p)}$ = –48.5 kcal mol^{-1}) and an unfavorable entropy of binding ($-T\Delta S^{s}_{w,P(r_p)}$= 2.2 kcal mol^{-1}). This buried hydration site is coordinated by two nearby water molecules, a threonine's carbonyl oxygen and a glutamate carboxylate. Further electrostatic stabilization is provided by a closely placed aspartate. Motions in this hydration site are highly restricted, hence the unfavorable entropy of binding.

Figure 10B depicts a different situation where water in the hydration site is more solvent accessible and connects to bulk. Water at this hydration site interacts with the amide side chain nitrogen atom as well as the amide backbone nitrogen of a glutamine. However, interactions with the carbonyl oxygens of a neighboring histidine and aspartate are also possible. Although water in this hydration site is hindered in its translations, it is able to form hydrogen-bonds in many ways with backbone donor/acceptors and neighboring bulk solvent. Consequently, the entropy of binding is more negative than in bulk water ($-T\Delta S^{s}_{w,P(r_p)}$= –0.7 kcal mol^{-1}).

Next, the components of the enthalpy and entropy of binding were investigated. Figure 11A shows the enthalpies of binding $\Delta H^{s}_{w,P(r_p)}$, the water-water component of the enthalpy of binding ΔH^{s}_{w} and the water-solute component of the enthalpy of binding $\Delta H^{s}_{P(r_p)}$. Evaluation of the distributions shows that the water-water enthalpic component tends to be unfavorable, whereas the water-solute enthalpic component is favorable. In more detail water-water enthalpies are above zero for 28.4% of the sites, whereas this never happens with water-solute enthalpies.

Figure 11B depicts the components of the entropy of binding. From the plot the percentage of sites in which components contribute favorably or unfavorably to the free energy can be estimated. In this dataset 86.2% of the orientational entropy components are thermodynamically unfavorable, and this is followed by librational entropy (80.4% unfavorable) and vibrational entropy (73.8%).

Further insights may be gained by evaluating correlations between the distributions shown in Figure 12. Entropy loss in the protein hydration layer has a maximum of 2.3 kcal mol^{-1}, slightly higher than the experimental limits of 2 kcal mol^{-1} suggested by Dunitz.[56] Figure 12A and 12B show that changes in orientational entropy correlate little with changes in vibrational or librational entropies. Figure 12C shows that there is stronger correlation between changes in vibrational and librational entropy, and the magnitude of changes in vibrational entropy are slightly larger. Finally the correlation between water-solute and water-water components of the enthalpy of binding is shown in Figure 12D. This indicates that water molecules that interact strongly with a protein (low $\Delta H^{s}_{P(r_p)}$values) also tend to interact strongly with neighboring water molecules (low ΔH^{s}_{w} values).

Conclusions

Extensive analysis of hydration sites surrounding 17 proteins has afforded a number of novel insights into binding thermodynamics at protein interfaces. On average water free energies are more negative near acidic amino-acids groups, followed by basic amino-acids, but differences between polar and non-polar amino acids are small. Differences in free energy distributions around each amino-acid were also evaluated, and revealed a broadly similar picture to the analysis of average binding free energies. Qualitatively, these results are similar to those reported by Beuming et al.[14] that used an IFST implementation for their analyses. However,

there is significant variability between the two methods in terms of the magnitude of the computed water enthalpies and entropies. These differences are attributed to the protocol used to define hydration sites, and the different theories used to calculate entropy. This discrepancy in computed water enthalpies of binding also is specific to the work by Beuming et al.[14] ranges found in other work described before have ranges of similar values.[50,51]

A comparison of density-clustered hydration sites and hydration sites resolved in X-ray diffracted protein structures reveals that the molecular simulations do tend to assign high-density hydration sites near X-ray resolved sites, but detect also many other high and low-density sites. Comparisons of computed enthalpies with Poisson-Boltzmann electrostatic calculations suggest that the stability of a water molecule may not be generally inferred from inspection of the magnitude of the local electrostatic potential. Rather, it is the nature of the coordination environment that must be taken into account. The thermodynamic properties of water in known ligand binding sites do not differ from the thermodynamic properties computed in other pockets. However, since binding sites tend to be made of the largest pocket, they can also be identified by evaluation of the water-accessible volume of a pocket. Lastly, high-density hydration sites are stabilized mostly by the enthalpy of interactions between the protein and water, and in rare occasions entropically stabilized by favorable changes in vibrational and librational entropy. The maximum energetic contribution of an entropy loss of a hydration site at a protein surface to the free energy is around +2.5 kcal mol^{-1} which correlates well with the estimate of +2 kcal mol^{-1} originally put forward by Dunitz.[56]

Future work in this topic could focus on parameterizing simple empirical models that may predict the molecular simulation computed water thermodynamics from rapid structural analysis of a protein. It would also be interesting to repeat similar analyses using more elaborate

definitions of the orientational entropy term such as those recently proposed by Henchman and coworkers.[57] Finally, as the protein structures studied here were rigid, it would be intriguing to explore how fluctuations in local binding thermodynamics are coupled to protein conformational changes.

ASSOCIATED CONTENT

Supporting Information. Heatmap of significance values of differences in amino acid water binding free energy distributions; tables of average water binding free energies of amino acids, and amino acid groups, distributions of water binding free energy distributions for each amino acid, distributions of amino acids around hydration sites with density cutoff 10× greater than bulk. This material is available free of charge via the Internet at http://pubs.acs.org

Acknowledgments

JM is supported by a University Research Fellowship from the Royal Society. RHH is supported by BBSRC Grant BB/K001558/1. This research was also supported by EPSRC through an award of a CASE studentship to GG. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Corresponding Author

* mail@julienmichel.net

Present Addresses

Max Planck Institute for Coal Research, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr

REFERENCES

(1) England, J. L.; Pande, V. S. Charge, Hydrophobicity, and Confined Water: Putting Past Simulations into a Simple Theoretical framework. *Biochem. Cell Biol.* **2010**, *88*, 359–369.
(2) Nicholls, A.; Sharp, K. A.; Honig, B. Protein Folding and Association: Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons. *Proteins Struct. Funct. Bioinforma.* **1991**, *11*, 281–296.
(3) Khandelwal, G.; Jayaram, B. DNA–Water Interactions Distinguish Messenger RNA Genes from Transfer RNA Genes. *J. Am. Chem. Soc.* **2012**, *134*, 8814–8816.
(4) Spyrakis, F.; Cozzini, P.; Bertoli, C.; Marabotti, A.; Kellogg, G. E.; Mozzarelli, A. Energetics of the Protein-DNA-Water Interaction. *BMC Struct. Biol.* **2007**, *7*, 1-18.
(5) Reddy, C. K.; Das, A.; Jayaram, B. Do Water Molecules Mediate Protein-DNA recognition?1. *J. Mol. Biol.* **2001**, *314*, 619–632.
(6) Conte, L. L.; Chothia, C.; Janin, J. The Atomic Structure of Protein-Protein Recognition sites1. *J. Mol. Biol.* **1999**, *285*, 2177–2198.
(7) Michel, J. Current and Emerging Opportunities for Molecular Simulations in Structure-Based Drug Design. *Phys. Chem. Chem. Phys.*, **2014**, *16*, 4465-4477.
(8) Jungwirth, P. Biological Water or Rather Water in Biology? *J. Phys. Chem. Lett.* **2015**, *6*, 2449–2451.
(9) Sebastiani, F.; Orecchini, A.; Paciaroni, A.; Jasnin, M.; Zaccai, G.; Moulin, M.; Haertlein, M.; De Francesco, A.; Petrillo, C.; Sacchetti, F. Collective THz Dynamics in Living Escherichia Coli Cells. *Chem. Phys.* **2013**, *424*, 84–88.
(10) Fogarty, A. C.; Laage, D. Water Dynamics in Protein Hydration Shells: The Molecular Origins of the Dynamical Perturbation. *J. Phys. Chem. B* **2014**, *118*, 7715–7729.
(11) Martin, D. R.; Matyushov, D. V. Hydration Shells of Proteins Probed by Depolarized Light Scattering and Dielectric Spectroscopy: Orientational Structure Is Significant, Positional Structure Is Not. *J. Chem. Phys.* **2014**, *141*, 22D501.
(12) Conti Nibali, V.; Havenith, M. New Insights into the Role of Water in Biological Function: Studying Solvated Biomolecules Using Terahertz Absorption Spectroscopy in Conjunction with Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2014**, *136*, 12800–12807.
(13) Persson, E.; Halle, B. Cell Water Dynamics on Multiple Time Scales. *Proc. Natl. Acad. Sci.* **2008**, *105*, 6266–6271.
(14) Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. Thermodynamic Analysis of Water Molecules at the Surface of Proteins and Applications to Binding Site Prediction and Characterization. *Proteins Struct. Funct. Bioinforma.* **2012**, *80*, 871–883.
(15) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein–ligand Binding. *Proc. Natl. Acad. Sci.* **2007**, *104*, 808–813.

(16) Henchman, R. H. Free Energy of Liquid Water from a Computer Simulation via Cell Theory. *J. Chem. Phys.* **2007**, *126*, 064504.
(17) Gerogiokas, G.; Calabro, G.; Henchman, R. H.; Southey, M. W. Y.; Law, R. J.; Michel, J. Prediction of Small Molecule Hydration Thermodynamics with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 35–48.
(18) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. Evaluation of Host–Guest Binding Thermodynamics of Model Cavities with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 4055–4068.
(19) Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Hefeitz, A.; Bodkin, M.; Law, R. J.; Michel, J. Evaluation of Water Displacement Energetics in Protein Binding Sites with Grid Cell Theory. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8416–8426.
(20) Woods, C.; Michel, J. *Sire Molecular Simulation Framework, Revision 1786*, http://siremol.org (accessed September 18, 2016).
(21) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.
(22) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584–591.
(23) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; McDonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Iyashiro, J. M.; Penning, T. D.; Seibert, K.; et al. Structural Basis for Selective Inhibition of Cyclooxygenase-2 by Anti-Inflammatory Agents. *Nature* **1996**, *384*, 644–648.
(24) Okamoto, Y.; Anan, H.; Nakai, E.; Morihira, K.; Yonetoku, Y.; Kurihara, H.; Sakashita, H.; Terai, Y.; Takeuchi, M.; Shibanuma, T.; et al. Peptide Based Interleukin-1 Beta Converting Enzyme (ICE) Inhibitors: Synthesis, Structure Activity Relationships and Crystallographic Study of the ICE-Inhibitor Complex. *Chem. Pharm. Bull. (Tokyo)* **1999**, *47*, 11–21.
(25) Arris, C. E.; Boyle, F. T.; Calvert, A. H.; Curtin, N. J.; Endicott, J. A.; Garman, E. F.; Gibson, A. E.; Golding, B. T.; Grant, S.; Griffin, R. J.; et al. Identification of Novel Purine and Pyrimidine Cyclin-Dependent Kinase Inhibitors with Distinct Molecular Interactions and Tumor Cell Growth Inhibition Profiles. *J. Med. Chem.* **2000**, *43*, 2797–2804.
(26) Podust, L. M.; Poulos, T. L.; Waterman, M. R. Crystal Structure of Cytochrome P450 14α-Sterol Demethylase (CYP51) from Mycobacterium Tuberculosis in Complex with Azole Inhibitors. *Proc. Natl. Acad. Sci.* **2001**, *98*, 3068–3073.
(27) Dvir, H.; Wong, D. M.; Harel, M.; Barril, X.; Orozco, M.; Luque, F. J.; Muñoz-Torrero, D.; Camps, P.; Rosenberry, T. L.; Silman, I.; et al. 3D Structure of Torpedo Californica Acetylcholinesterase Complexed with Huprine X at 2.1 Å Resolution: Kinetic and Molecular Dynamic Correlates,. *Biochemistry (Mosc.)* **2002**, *41*, 2970–2981.
(28) Maignan, S.; Guilloteau, J.-P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal Structures of Human Factor Xa Complexed with Potent Inhibitors. *J. Med. Chem.* **2000**, *43*, 3226–3232.
(29) Istvan, E. S.; Deisenhofer, J. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. *Science* **2001**, *292*, 1160–1164.

(30) Ala, P. J.; DeLoskey, R. J.; Huston, E. E.; Jadhav, P. K.; Lam, P. Y. S.; Eyermann, C. J.; Hodge, C. N.; Schadt, M. C.; Lewandowski, F. A.; Weber, P. C.; et al. Molecular Recognition of Cyclic Urea HIV-1 Protease Inhibitors. *J. Biol. Chem.* **1998**, *273*, 12325–12331.

(31) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal Structures of the Kinase Domain of c-Abl in Complex with the Small Molecule Inhibitors PD173955 and Imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236–4243.

(32) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; et al. Inhibition of p38 MAP Kinase by Utilizing a Novel Allosteric Binding Site. *Nat. Struct. Mol. Biol.* **2002**, *9*, 268–272.

(33) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the Epidermal Growth Factor Receptor Kinase Domain Alone and in Complex with a 4-Anilinoquinazoline Inhibitor. *J. Biol. Chem.* **2002**, *277*, 46265–46272.

(34) Marquis, R. W.; Ru, Y.; LoCastro, S. M.; Zeng, J.; Yamashita, D. S.; Oh, H.-J.; Erhard, K. F.; Davis, L. D.; Tomaszek, T. A.; Tew, D.; et al. Azepanone-Based Inhibitors of Human and Rat Cathepsin K. *J. Med. Chem.* **2001**, *44*, 1380–1395.

(35) Huai, Q.; Wang, H.; Sun, Y.; Kim, H.-Y.; Liu, Y.; Ke, H. Three-Dimensional Structures of PDE4D in Complex with Roliprams and Implication on Inhibitor Selectivity. *Structure* **2003**, *11*, 865–873.

(36) Puius, Y. A.; Zhao, Y.; Sullivan, M.; Lawrence, D. S.; Almo, S. C.; Zhang, Z.-Y. Identification of a Second Aryl Phosphate-Binding Site in Protein-Tyrosine Phosphatase 1B: A Paradigm for Inhibitor Design. *Proc. Natl. Acad. Sci.* **1997**, *94*, 13420–13425.

(37) Gordon, E.; Mouz, N.; Duée, E.; Dideberg, O. The Crystal Structure of the Penicillin-Binding Protein 2x from Streptococcus Pneumoniae and Its Acyl-Enzyme Form: Implication in Drug resistance1. *J. Mol. Biol.* **2000**, *299*, 477–485.

(38) Sung, B.-J.; Yeon Hwang, K.; Ho Jeon, Y.; Lee, J. I.; Heo, Y.-S.; Hwan Kim, J.; Moon, J.; Min Yoon, J.; Hyun, Y.-L.; Kim, E.; et al. Structure of the Catalytic Domain of Human Phosphodiesterase 5 with Bound Drug Molecules. *Nature* **2003**, *425*, 98–102.

(39) Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; al, et. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **1996**, *274*, 948–953.

(40) D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R.; Luo, R.C. Walker, W. Zhang, K.M. Merz, B. et al. *AMBER 11*; University of California, San Francisco, 2010.

(41) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *65*, 712–725.

(42) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665–9678.

(43) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

(44) Tironi, I. G.; Sperb, R.; Smith, P. E.; Gunsteren, W. F. van. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102* (13), 5451–5459.
(45) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure And/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.
(46) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
(47) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168.
(48) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884–1897.
(49) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proc. Natl. Acad. Sci.* **2001**, *98*, 10037–10041.
(50) Huggins, D. J. Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations. *Biophys. J.* **2015**, *108*, 928–936.
(51) Li, Z.; Lazaridis, T. Water at Biomolecular Binding Interfaces. *Phys. Chem. Chem. Phys.* **2007**, *9*, 573–581.
(52) Hamelberg, D.; McCammon, J. A. Standard Free Energy of Releasing a Localized Water Molecule from the Binding Pockets of Proteins: Double-Decoupling Method. *J. Am. Chem. Soc.* **2004**, *126*, 7683–7689.
(53) Michel, J.; Tirado-Rives, J.; Jorgensen W. L. Prediction of the water content in protein binding sites. *J. Phys. Chem. B* **2009**, *113*, 13337-13346.
(54) Michel, J.; Tirado-Rives, J.; Jorgensen W. L. Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.* **2009**, *131*, 15403-15411.
(55) Vukovic, S.; Brennan, P. E.; Huggins, D. J. Exploring the Role of Water in Molecular Recognition: Predicting Protein Ligandability Using a Combinatorial Search of Surface Hydration Sites. *J. Phys. Condens. Matter* **2016**, *28*, 344007.
(56) Dunitz, J. D. The Entropic Cost of Bound Water in Crystals and Biomolecules. *Science* **1994**, *264*, 670–670.
(57) Henchman, R. H.; Cockram, S. J. Water's Non-Tetrahedral Side. *Faraday Discuss.* **2013**, *167*, 529.

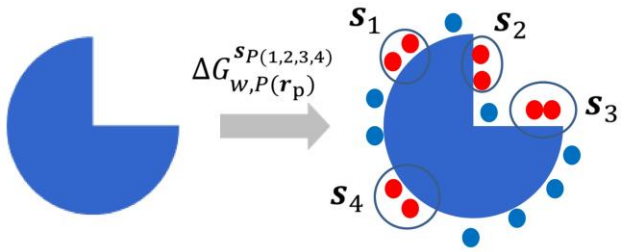

Figure 1. Evaluation of binding energies of a region **s,** typically in the vicinity of residues or pockets of a protein *P*. Proteins are depicted by large blue spheres. In all GCT analyses, water molecules (red circles) inside the monitored regions, $\mathbf{s}_{P(1,2...n)}$, contribute to the computed binding free energies, whereas those that are out of the monitored regions (in blue) are not considered. The subscript $\mathbf{r}_p$ indicates that the protein coordinates were restrained during the analysis.

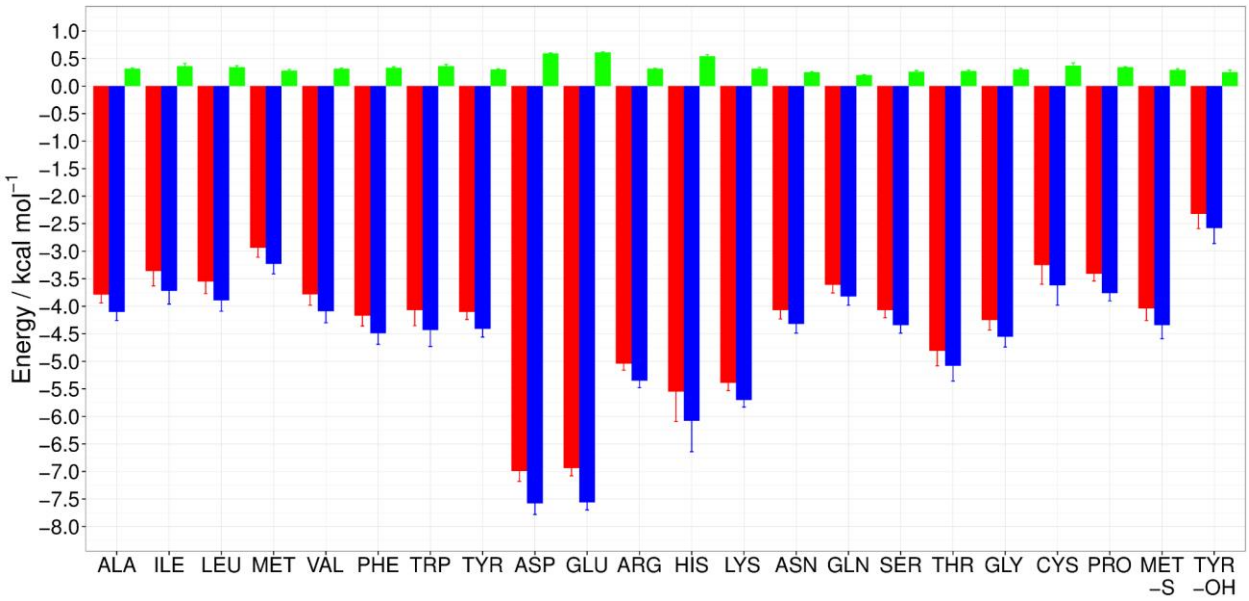

Figure 2. The average values of $\Delta G^{s,w}_{w,P(r_p)}$ (red), $\Delta H^{s,w}_{w,P(r_p)}$ (blue), $-T\Delta S^{s,w}_{w,P(r_p)}$ (green) around all the amino acids. The error bars represent the standard error of the mean.All plots were generated with the ggplot2 package of R unless stated otherwise.[48]

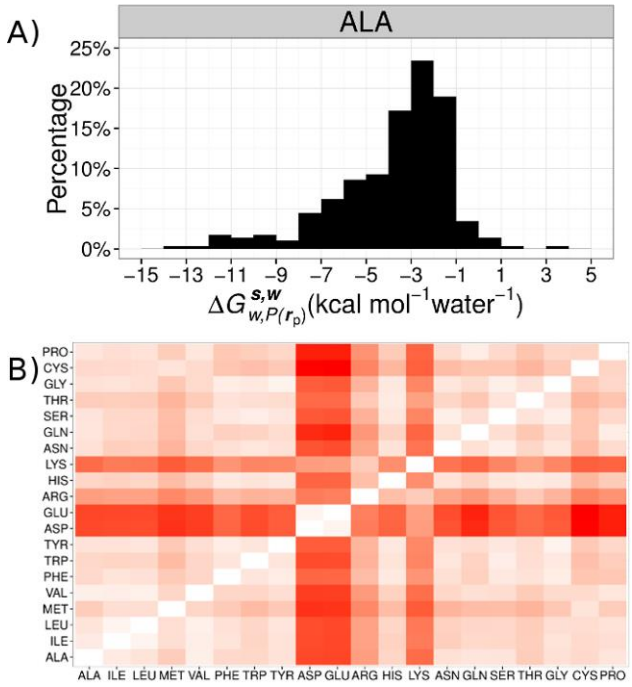

Figure 3. A) One example of an empirical distribution of water free energies around alanine side-chains. B) Heatmap of Kolmogorov-Smirnov *D* statistics between empirical cumulative per-water $\Delta G^{s,w}_{w,P(r_p)}$ distribution functions. *D* values range from 0 (white) to 0.7415 (red).

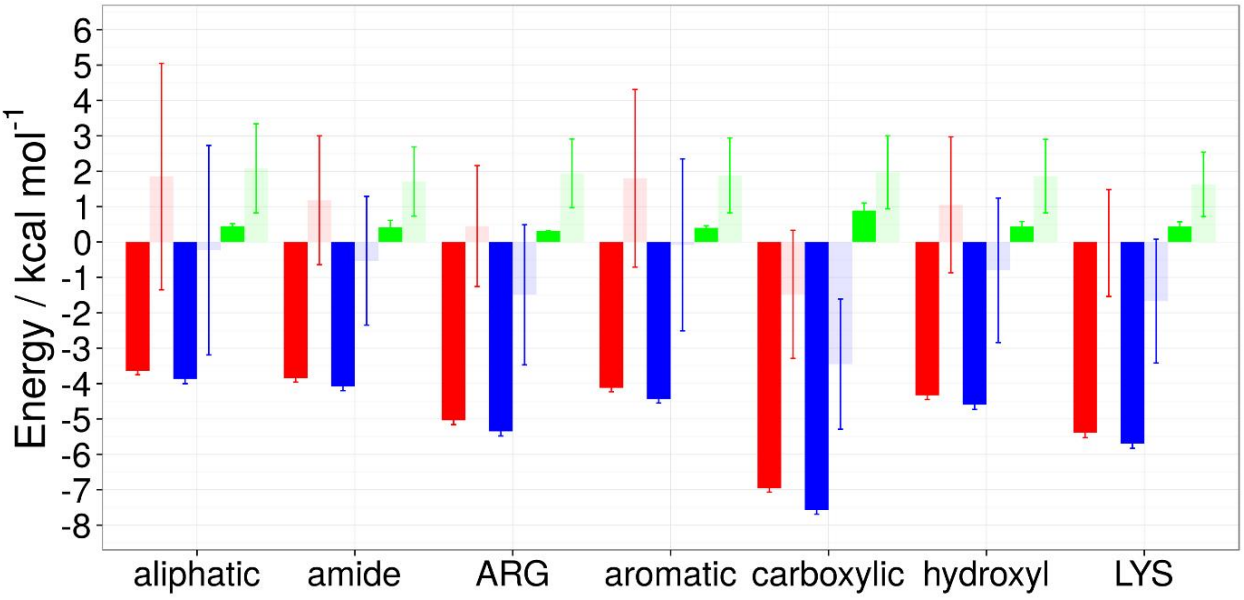

Figure 4. The average values of $\Delta G^{s,w}_{w,P(r_p)}$ (red), $\Delta H^{s,w}_{w,P(r_p)}$ (blue), $-T\Delta S^{s,w}_{w,P(r_p)}$ (green) around groups of amino acids. The shaded bars correspond to the IFST results of Beuming et al.[49] For the GCT results the error bars denote the standard error of the mean.

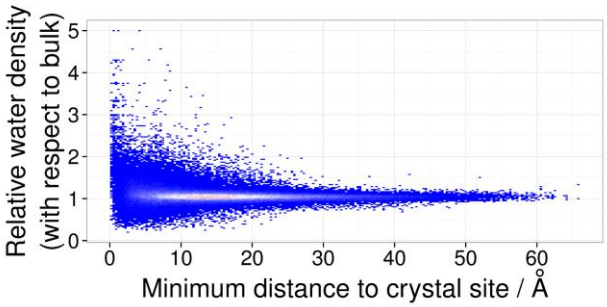

Figure 5. Two-dimensional probability distribution of hydration sites. The *x* axis measure the minimum distance to a hydration site observed in a X-ray diffracted protein structure. The *y* axis measures the density of the site relative to bulk. Probabilities are coloured from low (blue) to high (red).

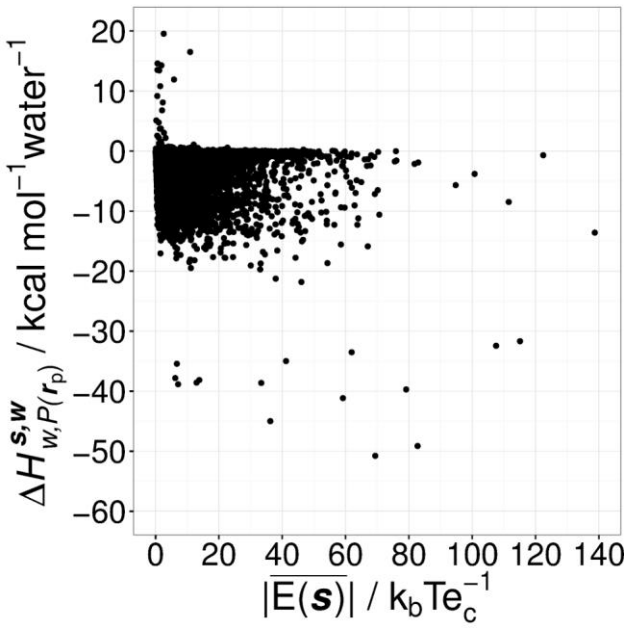

Figure 6. Correlation between the average magnitude of the electrostatic potential and the GCT computed binding enthalpies of hydration sites per-water, $\Delta H^{s,w}_{w,P(r_p)}$.

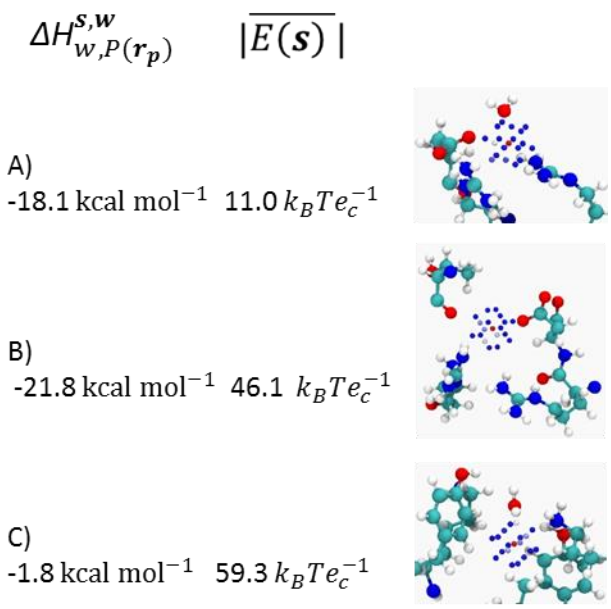

Figure 7. Selected hydration sites differing considerably in the magnitude of the average electrostatic potential and the enthalpy of binding. These sites were obtained from a simulation of PDB structure 1E1X (cyclin-dependent kinase 2). Panels A), B) and C) denote various cases were the magnitude of the local electrostatic potential correlates is compared with the enthalpy of hydration of the site. Grid points related to the centroid are colored from low relative water density to high relative water density using a color range from blue-white-red. For A) the range varies from 0-16.4, B) 0-8.1 and C) 0-12.5 relative water density.

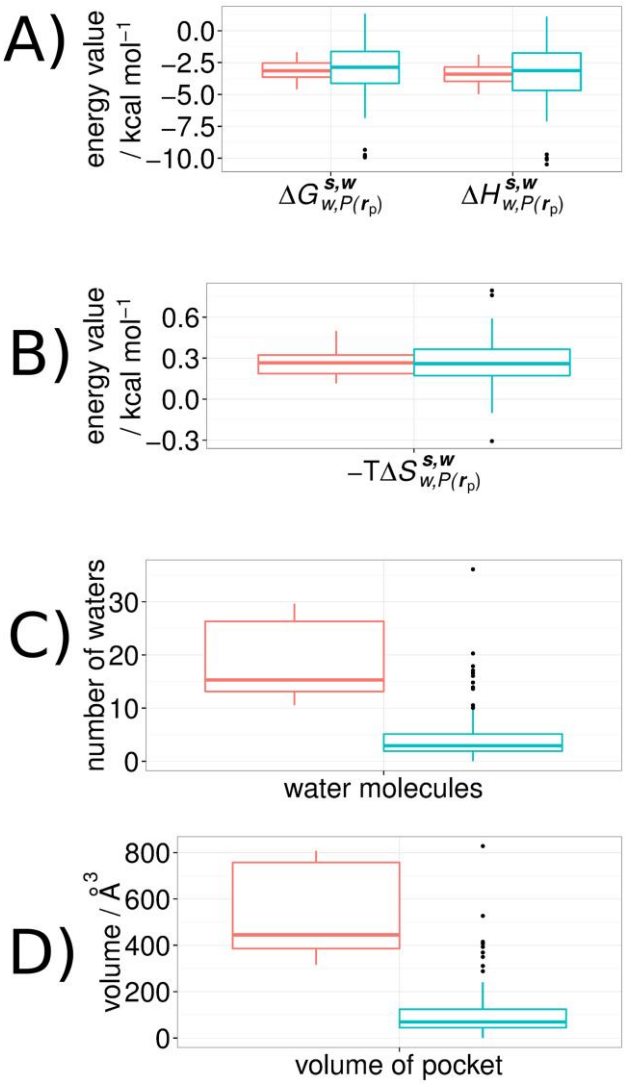

Figure 8. Boxplot comparison of binding sites (red) and pockets (blue) properties. The box plots show the median and the upper and lower quartile of the distributions of per-water properties. A) Free energy and enthalpy of binding. B) Entropy of binding. C) Distributions of the number of water molecules and D) the volumes of the pockets. Outliers outside 1.5 × the interquartile range are shown as dots.

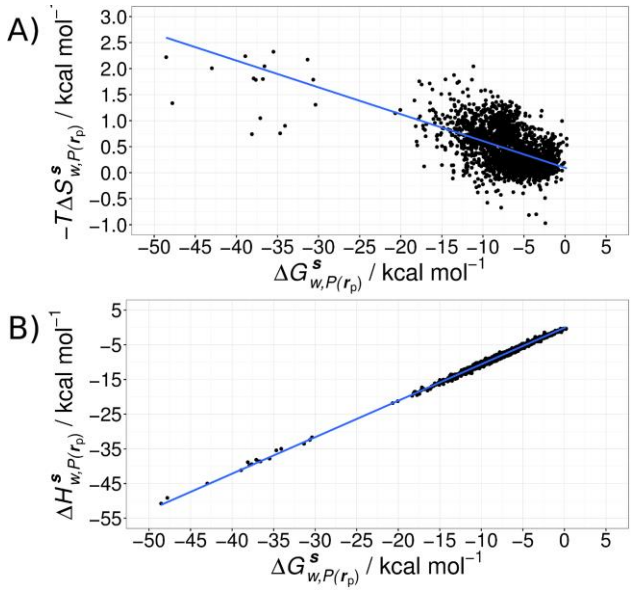

Figure 9. Correlation of thermodynamic components for high-density hydration sites. A) Correlation of $\Delta G^{\mathbf{s}}_{w,P(\boldsymbol{r}_p)}$ with $-\mathrm{T}\Delta S^{\mathbf{s}}_{w,P(\boldsymbol{r}_p)}$. B) Correlation of $\Delta G^{\mathbf{s}}_{w,P(\boldsymbol{r}_p)}$ with $\Delta H^{\mathbf{s}}_{w,P(\boldsymbol{r}_p)}$.

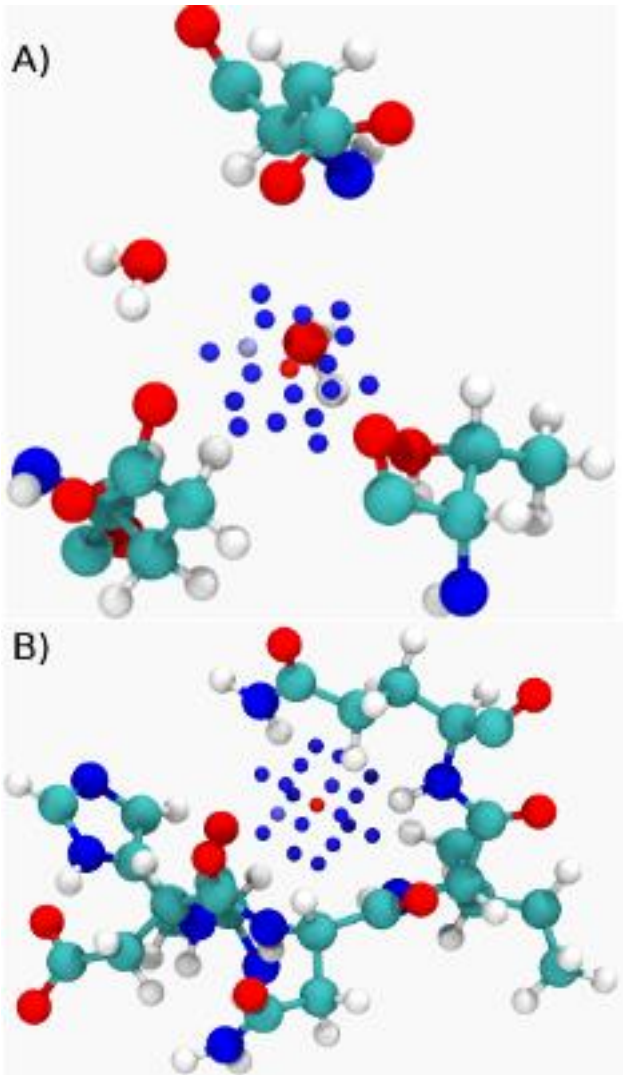

Figure 10. Selected hydration sites with unusual entropies of binding. A) Hydration site taken from the simulation of 1OYN. $\Delta G^{s}_{w,P(r_p)}$ is -48.5 kcal mol^{-1} and $-T\Delta S^{s}_{w,P(r_p)}$ is +2.2 kcal mol^{-1}. B) Hydration site taken from the simulation of 1E66 simulation. $\Delta G^{s}_{w,P(r_p)}$ is -7.8 kcal mol^{-1} and $-T\Delta S^{s}_{w,P(r_p)}$ is -0.7 kcal mol^{-1}. Grid points are color-coded by water density from low (blue) to high (red).

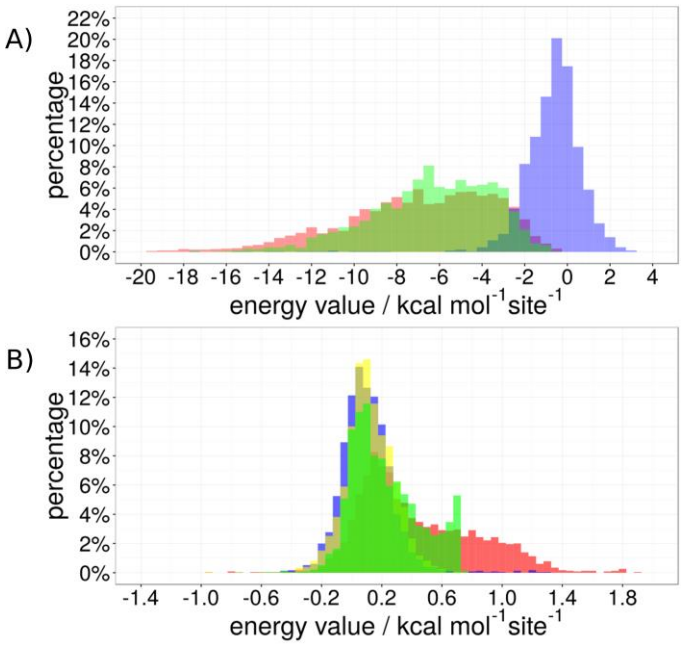

Figure 11. A) Probability distribution of the components of the $\Delta H^{s}_{w,P(\boldsymbol{r}_p)}$ (red), ΔH^{s}_{w} (blue) and $\Delta H^{s}_{P(\boldsymbol{r}_p)}$ (green). The water-solute term has a long tail that extends below the left hand side of the *x*-axis. B) Probability distribution of the components of the entropy of binding (red), $-T\Delta S^{s,ori}_{w,P(\boldsymbol{r}_p)}$ (green), $-T\Delta S^{s,lib}_{w,P(\boldsymbol{r}_p)}$ (orange) and $-T\Delta S^{s,vib}_{w,P(\boldsymbol{r}_p)}$ (blue).

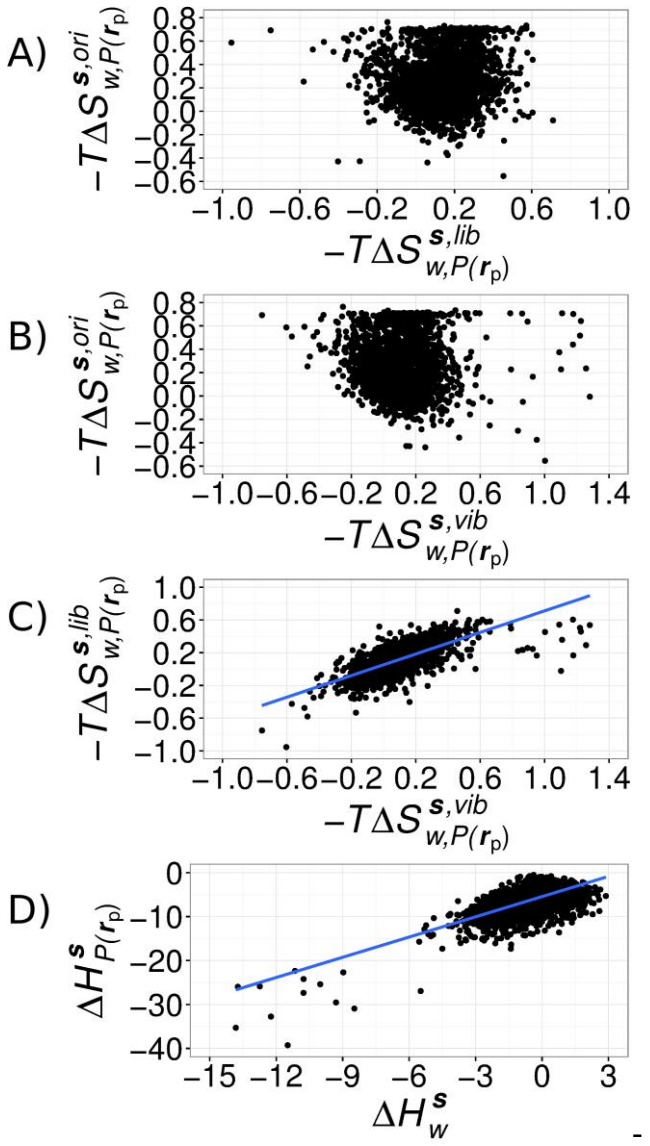

Figure 12. Correlation plots between A) $-T\Delta S^{s,ori}_{w,P(r_p)}$ and $-T\Delta S^{s,lib}_{w,P(r_p)}$, B) $-T\Delta S^{s,ori}_{w,P(r_p)}$ and $-T\Delta S^{s,vib}_{w,P(r_p)}$ C) $-T\Delta S^{s,lib}_{w,P(r_p)}$ and $-T\Delta S^{s,vib}_{w,P(r_p)}$ and D) $\Delta H^{s}_{P(r_p)}$ and $\Delta H^{s}_{w}$with all values in kcal mol^{-1}.

Table of Contents Graphic